Lesson 19: Comparing Data Distributions

As you have seen in previous lessons, it can be difficult to understand a data set just by looking at raw data. Often, readers want to have a concise and useful summary.

This becomes extremely important when data distributions are compared to one another. While a reader may be interested in knowing if a typical adult male polar bear in Alaska is larger than a typical adult male grizzly bear in British Columbia, it would also be useful to be able to compare the variability and shape of the distributions of these two groups of bears as well. With summary graphs of the two distributions placed side-by-side, you can more easily assess and compare the characteristics of one distribution to the other distribution.

By this point, you should have completed the collection of data for your statistical question. This lesson will provide graphical representations of data distributions that are part of the summaries expected in your project.

Classwork

Example 1: Comparing Groups Using Box Plots

Recall that a *box plot* is a visual representation of a 5-number summary. It is drawn with careful reference to a number line, so the difference between any two values in the 5-number summary is represented visually as a distance. For example, the box of a box plot is drawn so that width of the box represents the IQR. The whiskers (the lines that extend from the box) are drawn such that the distance from the far end of one whisker to the far end of the other whisker represents the range. If two box plots (each representing a different distribution) were drawn side-by-side using the same scale, one could quickly compare the IQRs and ranges of the two distributions while also gaining a sense of the 5-number summary values for each distribution.

Here is a data set of the ages of 43 participants in a local 5-kilometer race (shown in a previous lesson).

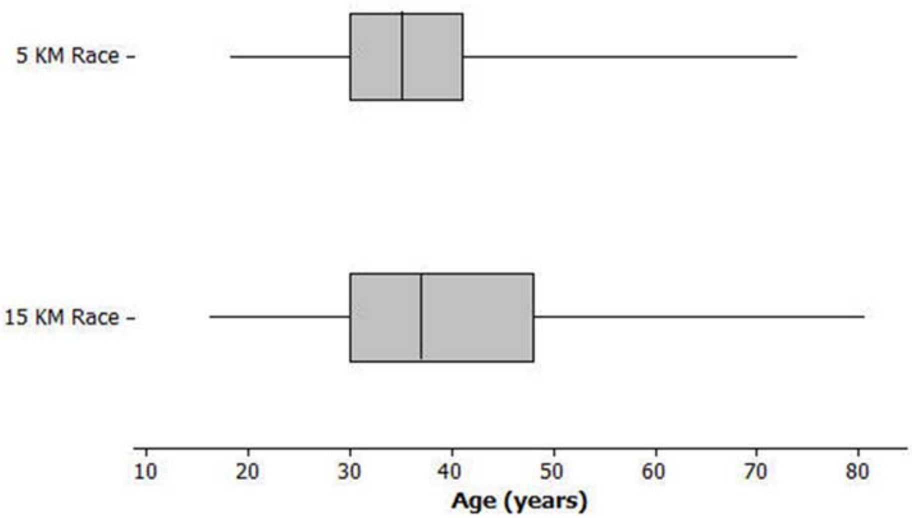
20	30	30	35	36	34	38	46
45	18	43	23	47	27	21	30
32	32	31	32	36	74	41	41
51	61	50	34	34	34	35	28
57	26	29	49	41	36	37	41
38	30	30					

Here is the 5-number summary for the data: Minimum = 18, Q1 = 30, Median = 35, Q3 = 41, Maximum = 74.

Later that year, the same town also held a 15-kilometer race. The ages of the 55 participants in that race appear below.

47	19	30	30	36	37	35	39
19	49	47	16	45	22	50	27
19	20	30	32	32	31	32	37
22	81	43	43	54	66	53	35
22	35	35	36	28	61	26	29
38	52	43	37	38	43	39	30
58	30	48	49	54	56	58	

Does the longer race appear to attract different runners in terms of age? Here are side-by-side box plots that may help answer that question. Side-by-side box plots are two or more box plots drawn



Exercises 1–6

1. Based on the side-by-side box plots, estimate the 5-number summary for the 15-kilometer race data set.

2. Do the two data sets have the same median? If not, which race had the higher median age?

3. Do the two data sets have the same IQR? If not, which distribution has the greater spread in the middle 50% of its distribution?

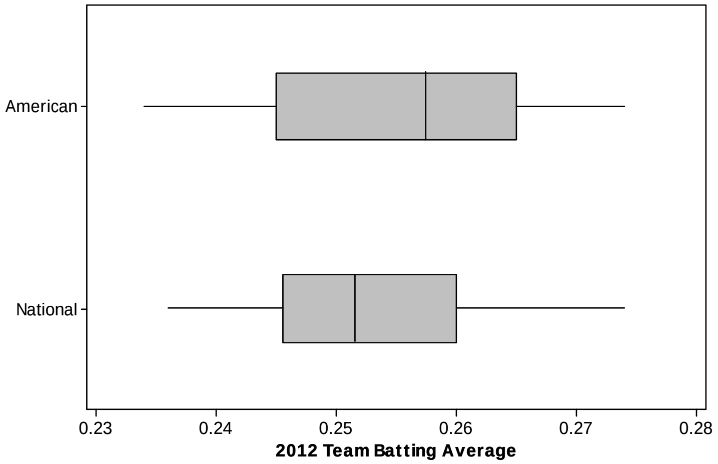
4. Which race had the smaller overall range of ages? What do you think the range of ages is for the 15-kilometer race?

5. Which race had the oldest participant? About how old was this participant?

6. Now consider just the youngest 25% of participants in the 15-kilometer race. How old was the youngest runner in this group? How old was the oldest runner in this group? How does that compare with the 5-kilometer race?

Exercises 7–12: Comparing Box Plots

In 2012, Major League Baseball was comprised of two leagues: an American League of 14 teams and a National League of 16 teams. It is believed that the American League teams would generally have higher values of certain offensive statistics such as batting average and on-base percentage. (Teams want to have high values of these statistics.) Use the following side-by-side box plots to investigate these claims. (Source: http://mlb.mlb.com/stats/sortable.jsp accessed May 13, 2013)

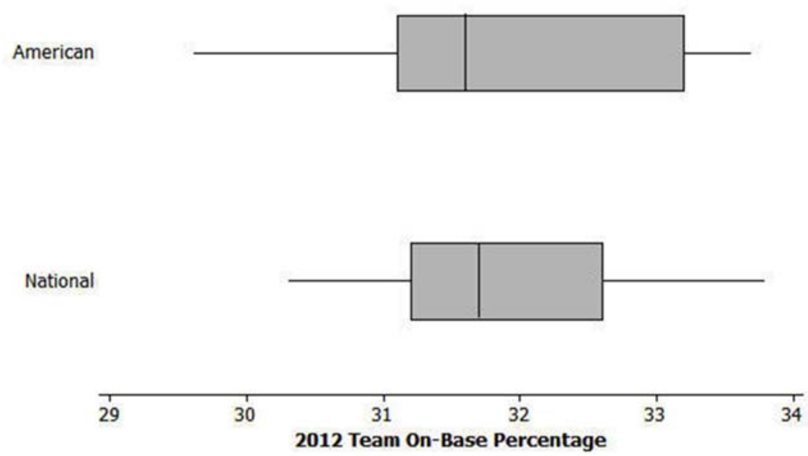


7. Was the highest American League team batting average very different from the highest National League team batting average? If so, approximately how large was the difference and which league had the higher maximum value?

8. Was the range of American League team batting averages very different or only slightly different from the range of National League team batting averages?

9. Which league had the higher median team batting average? Given the scale of the graph and the range of the data sets, does the difference between the median values for the two leagues seem to be small or large? Explain why you think it is small or large.

10. Based on the box plots below for on-base percentage, which 3 summary values (from the 5-number summary) appear to be the same or virtually the same for both leagues?



11. Which league's data set appears to have less variability? Explain.

12. Respond to the original statement: "It is believed that the American League teams would generally have higher values of … on-base percentage." Do you agree or disagree based on the graphs above? Explain.

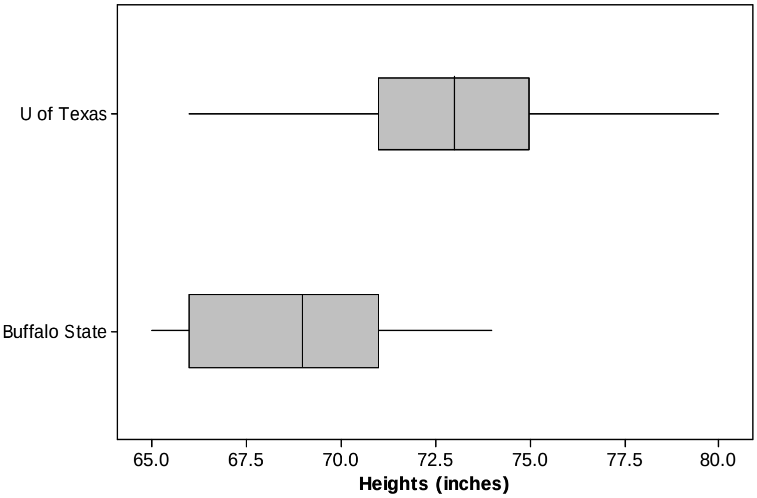
Lesson Summary

When comparing the distribution of a quantitative variable for two or more distinct groups, it is useful to display graphs of the groups' distributions side-by-side using the same scale. Generally, you can more easily notice, quantify, and describe the similarities and differences in the distributions of the groups.

Problem Set

1. College athletic programs are separated into divisions based on school size, available athletic scholarships, and other factors. A researcher is curious to know if members of swimming and diving programs in Division I (schools that offer athletic scholarships and tend to have large enrollment) are generally taller than the swimmers and divers in Division III programs (schools that do not offer athletic scholarships and tend to have smaller enrollment). To begin the investigation, the researcher creates side-by-side box plots for the heights (in inches) of members of the 2012–2013 University of Texas Men's Swimming and Diving Team (a Division I program) and the heights (in inches) of members of the 2012–2013 Buffalo State College Men's Swimming and Diving Team (a Division III program).

   (From http://www.texassports.com/sports/m-swim/mtt/tex-m-swim-mtt.html accessed April 30, 2013, all 41 member heights listed, and http://www.buffalostateathletics.com/roster.aspx?path=mswim& accessed May 15, 2013, 11 members on roster; only 10 heights were listed)



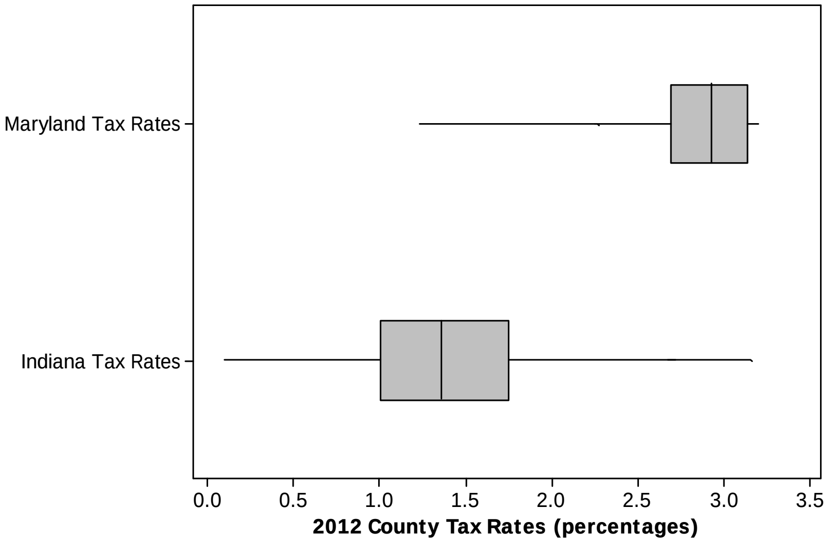
   a. Which data set has the smaller range?

   b. True or False: A team member of median height on the University of Texas team would be taller than a team member of median height on the Buffalo State College team.

   c. To be thorough, the researcher will examine many other college's sports programs to further investigate her claim that members of swimming and diving programs in Division I are generally taller than the swimmers and divers in Division III. But given the graph above, in this initial stage of her research, do you think that her claim might be valid? Carefully support your answer using comparative summary measures or graphical attributes.

2. Different states use different methods for determining a person's income tax. However, Maryland and Indiana both have systems where a person pays a different income tax rate based on the county in which he/she lives. Box plots summarizing the 24 different county tax rates for Maryland's 23 counties and Baltimore City (taxed like a county in this case) and the resident tax rates for 91 counties in Indiana in 2012 are shown below.

   (From http://taxes.marylandtaxes.com/Individual_Taxes/Individual_Tax_Types/Income_Tax/Tax_Information/Tax_Rates/Local_and_County_Tax_Rates.shtml accessed May 5, 2013 and www.in.gov/dor/files/12-county-rates.pdf accessed May 16, 2013)



   a. True or False: At least one Indiana county income tax rate is higher than the median county income tax rate in Maryland. Explain how you know.

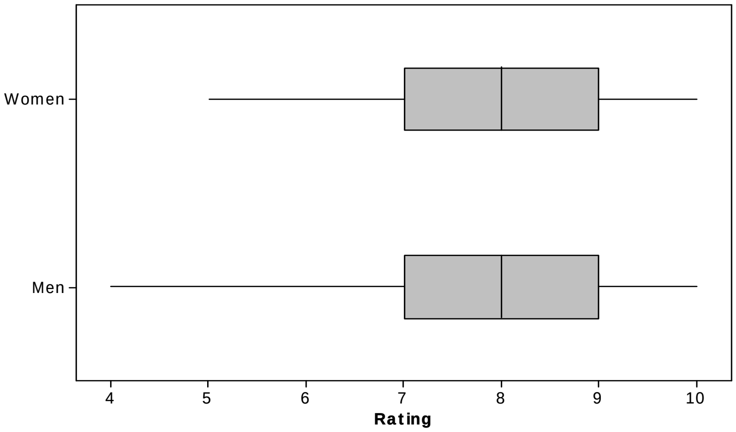
   b. True or False: The 24 Maryland county income tax rates have less variability than the 91 Indiana county income tax rates. Explain how you know.

   c. Which state appears to have typically lower county income tax rates? Explain.

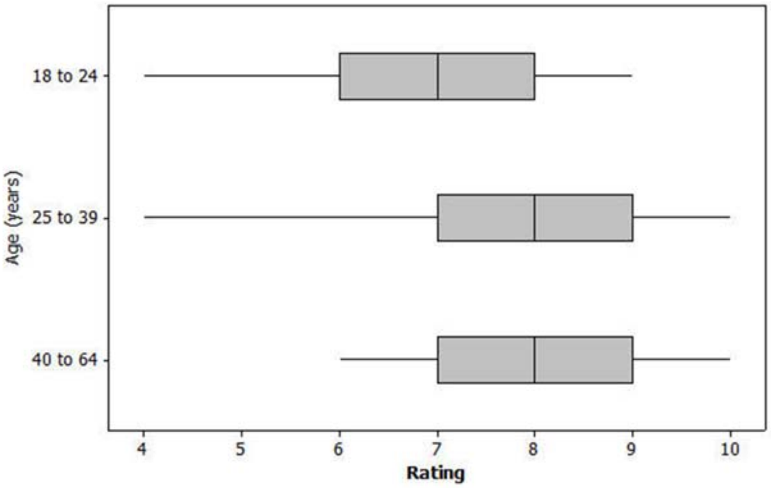
3. Many movie studios rely heavily on customer data in test markets to determine how a film will be marketed and distributed. Recently, previews of a soon to be released film were shown to 300 people. Each person was asked to rate the movie on a scale of 0 to 10, with 10 representing "best movie I've ever seen" and 0 representing "worst movie I've ever seen."

Below are some side-by-side box plots that summarize the ratings based on certain demographic characteristics.

For 150 women and 150 men:



For 3 distinct age groups:



a. Generally, does it appear that the men and women rated the film in a similar manner or in a very different manner? Write a few sentences explaining your answer using comparative information about center and spread from the graph.

b. Generally, it appears that the film typically received better ratings from the older members of the group. Write a few sentences using comparative measures of center and spread or graphical attributes to justify this claim.